A multi-factor approach to identifying missed synonymy in the UMLS

Bastien Rance, Olivier Bodenreider
National Library of Medicine

{bastien.rance, Olivier.bodenreider}@nih.gov

Motivation: In the Unified Medical Language System (UMLS), the eponymic term "Brodie abscess" denotes a chronic metaphyseal abscess of a bone, named after Sir Benjamin Collins Brodie, a 19th-Century British physician. Now, is "Brian abscess" - also a term of the UMLS - an abscess named after some Dr. Brian or a misspelled form of the term "Brain Abscess"? Similarly, the two terms "anti-diarrheals" (hyphenated) and "antidiarrheals" (one-word) seem to denote the same pharmacological class, yet are separate concepts in the UMLS. Misspellings and, more generally, uncontrolled lexical variation in biomedical terminologies lead to missed synonymy and have detrimental consequences on data retrieval, reasoning or literature mining processes.

Material and methods: The UMLS is a terminology integration system developed by the National Library of Medicine. The 2011AB version of the UMLS integrates medical concepts from 161 biomedical vocabularies and contains more than 2.6M concepts. The UMLS Metathesaurus groups into a single UMLS concept all synonymous names found in the source vocabularies for a given biomedical entity. The integration of the source vocabularies associates automatic lexical processing and human review. Despite the quality assurance procedures built in the UMLS development process, errors have been reported. In this study, we focus on the detection of missed synonymy, i.e. distinct UMLS concepts that have the same meaning and should be merged. We build a set of 1.5M missed synonymy candidates based on lexical similarity. We collect lexical, semantic and structural information in the UMLS, as well as contextual information from the biomedical literature and internet sources. We mine this information through a variety of techniques, including computation of similarity metrics, decision trees and vector space models, combined into a system for predicting missed synonymy.

Results: Our model has a precision of 0.71, and a recall of 0.74. Applied to the set of candidates, it identified 515 potential missed synonyms.

Conclusion: Because it has sufficient precision, our approach provides effective assistance to the Metathesaurus editors. Some of the errors detected have already been reported to and corrected by the developers of source terminologies. Although limited to lexically close candidates in this study, our approach can be used in a more general context.

Acknowledgments: This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).